УДК 004.93

Л.Г. Ахметшина, Т.С. Ямнич
Национальный горный университет, г. Днепропетровск, Украина
akhmlu@mail.ru

Интерполяция пространственных данных методом двумерного проецирования нечетких кластеров

В статье рассматриваются информационные возможности нового метода интерполяции пространственных данных, заданных на неравномерной сетке. Суть метода заключается в использовании нечеткой кластеризации, рассмотрении проекции функций принадлежности на координатную плоскость, вычислении степени принадлежности имеющихся данных и использовании их для получения прогнозных значений в дополнительных точках. Представлены результаты моделирования и реальной проверки работоспособности метода.

Введение

Существует множество задач, решение которых связано с обработкой неравномерно распределенных данных. В особенности это касается больших систем (геология, экология, горнодобывающая промышленность, экономика), для которых во многих случаях возможно измерение лишь в некоторых рабочих точках. Среди прочих следует выделить задачи обработки пространственных данных, при работе с которыми принципиальным вопросом является учет системы координат. Например, моделирование рельефа и изолиний земной поверхности, оценка границ возможных зон подтопления и загрязнения, оценка рудного тела, вычисление объемов под объектами на поверхности (при выемке угля и последующей закладке для предотвращения оседания земной поверхности) и т.д. Учитывая их практическую значимость, решению данной проблемы уделяется большое внимание.

Постановка задачи

Часто реальные экспериментальные данные, используемые для моделирования, являются неполными и/или неравномерно распределенными в пространстве, что сказывается на качестве результата. Различные алгоритмы интерполяции обеспечивают получение прогнозных значений в дополнительных точках на основе использования имеющихся данных [1]. При анализе пространственных данных следует учитывать их особенности, прежде всего свойства пространственной корреляции и топологии.

Условно можно выделить три основные группы методов интерполяции пространственных данных: *метод обратных взвешенных расстояний* (ОВР), *метод поверхности тренда и кригинг*. Именно они реализованы в большинстве современных геоинформационных (ГИС) и САПР системах, таких, как ArcGIS 9.0, Golden Software Surfer 8, MapInfo Professional.

В основе ОВР лежит понятие «окрестности» точки, влияющей на вклад рассчитываемых данных при определении интерполируемого значения. Вариации расчета расстояний между точками и способ их учета определяют различные модификации этого подхода.

Метод поверхности тренда направлен на выявление общих тенденций поверхности, а не точного моделирования мелких неровностей. Как и в предыдущем методе, используется начальный набор данных в пределах заданной окрестности, которая строится на основе любого метода со взвешиванием и определяется поверхность наилучшего приближения на основе математических уравнений, таких, как полиномы и сплайны.

В основе кригинга лежит учет статистических свойств поверхности и утверждение, что поверхность не может моделироваться только одним математическим уравнением. Рассматриваются три независимые величины: дрейф – представляет поверхность как общий тренд в определенном направлении, отклонения от общей тенденции – описывает случайные, но пространственно коррелированные незначительные поверхностные структуры, и случайный шум, который не имеет пространственной корреляции, каждая из которых обрабатывается отдельно.

При работе этих алгоритмов необходимо учитывать следующие факторы:

– область, в пределах которой производится интерполяция, должна быть окружена точками с известными значениями со всех сторон для возможности расширения окрестности по всем направлениям, иначе возникает вероятность появления ошибки вдоль границы;

– количество точек существенно влияет на точность модели, особенно при работе со сложными формами рельефа, но существует предел, при котором увеличение их приводит не только к значительному увеличению времени вычисления, но и к непредвиденным результатам;

– используемый метод расчета центра тяжести данных по полигону приводит в некоторых случаях к необходимости ручной корректировки хода вычислений.

Все описанные методы используют как базовое понятие «ближайшей» окрестности точки, которое по своей природе является неточным и может быть описано с использованием нечеткой логики.

Несмотря на то, что исследования в области решения задач с использованием нечеткого подхода ведутся достаточно широко практически во всех областях, прикладные результаты в области моделирования пространственной информации фактически отсутствуют.

Цель работы

В данной работе демонстрируется возможность использования нового подхода интерполяции пространственных данных, заданных на неравномерной сетке, методом двумерного проецирования нечетких кластеров, который в рамках рассматриваемой задачи позволяет частично устранить отмеченные недостатки традиционных методов.

Решение задачи

Возможность использования нечеткой логики для решения задач интерполяции была отмечена в [2]. Основное преимущество нечетких моделей, по сравнению с традиционными математическими моделями, связано с возможностью использования для их разработки значительно меньших объемов информации о системе, причем она может носить приближенный характер.

Нечеткая модель задает интерполяционную поверхность между точками пространства входов X и выходов Y, задаваемыми с помощью логических правил, использующих функцию принадлежности. Функция принадлежности ставит в соответствие

каждому значению x заданной переменной некоторое число из интервала $[0,1]$: $\mu_A(x): X \to [0,1], \forall x \in X$, характеризующее степень принадлежности элемента x к нечеткому множеству A. Каждое правило определяет важную типовую особенность поведения системы, геометрически соответствующую «существенной» точке пространства $X \times Y$, которая не всегда располагается непосредственно на характеристике реальной системы. Существуют различные походы определения «существенных» точек. В предлагаемом методе их месторасположение вычислялось с использованием метода нечеткой кластеризации (FCM).

В результате работы алгоритма FCM каждой точке входных данных ставится в соответствие вектор из функций принадлежности к каждому классу, на основе которого можно делать выводы о природе данного объекта. Задача нечеткой кластеризации формулируется следующим образом: на основе исходных данных D определить такое нечеткое разбиение $\Re(A) = \{A_k \mid A_k \subseteq A\}$ или нечеткое покрытие $\Im(A) = \{A_k \mid A_k \subseteq A\}$ множества А на заданное число c нечетких кластеров $A_k(k \in \{2,...,c\})$, которое доставляет экстремум некоторой целевой функции $f(\Re(A))$ среди всех нечетких разбиений или экстремум целевой функции $f(\Im(A))$ среди всех нечетких покрытий.

Основная идея метода двумерного проецирования нечетких кластеров заключается в предположении о том, что если образец принадлежит кластеру рассматриваемого класса в n-мерном пространстве, то его проекции также принадлежат проекциям этого кластера на любое из двумерных пространств $X_l \times X_r$, а функцию принадлежности заданного класса $\mu(x_1, x_2,..., x_n)$ можно определить через функции принадлежности его проекций на отдельные подпространства $\mu(x_1, x_2),...\mu(x_1, x_n),...\mu(x_{n-1}, x_n)$.

Алгоритм метода включает последовательность следующих действий.

1. *Определение центров кластеров с применением метода FCM*. На основе неравномерных экспериментальных измерений в пространстве признаков $X_1, X_2,..., X_n, Y$ (Y – целевой, интерполяция которого производится) осуществляется нечеткая кластеризация исходных данных.

Результатом являются значения центров кластеров $m_1, m_2,..., m_c$, вычисляемые итеративно по формуле

$$m_{xl}(t+1) = \frac{\sum_{j=1}^{N} \mu_{ij}^{q}(t) \cdot x_{lj}}{\sum_{j=1}^{N} \mu_{ij}^{q}(t)}, \quad (1)$$

где $(\forall i \in \{2,...,c\})$, c – количество классов; N – число объектов кластеризации, l – количество информативных признаков, причем $x_{n+1} = y$; q – параметр фаззификации, определяющий нечеткость кластера.

2. *Определение функций принадлежности кластеров*. Определение кластеров на шаге 1 осуществляется на всем пространстве $X \times Y$, а в практических задачах обычно требуется определить некоторое значение y при заданном значении входного вектора $\boldsymbol{x}$. Поэтому функции принадлежности следует формировать раздельно для входных и выходного параметров. Данная цель достигается построением проекций $m_i^{X_j}$

центров кластеров на координатное пространство, где $j = 1,...n$, $i = 1,...,c$, которые и рассматриваются в качестве «существенных» точек, являющихся центрами функций принадлежности μ_{ij}^X каждого из классов:

$$\mu_{ij}^X(x) = \frac{1}{\sum_{k=1}^{c}\left[\frac{d_{ij}^X(x)}{d_{kj}^X}\right]^{\frac{2}{q-1}}} = \mu_i^X(x_j), \tag{2}$$

где расстояние между вектором $\boldsymbol{x}_i$ и центром кластера m_i^X определяется только во входном пространстве по формуле:

$$d_{ij}^X(x) = \left|x - m_i^X\right| = \sqrt{\left(x_j - m_i^X\right)^T A\left(x_j - m_i^X\right)}, \tag{3}$$

размерность положительно определенной симметричной матрицы A равна $n \times n$.

3. *Вычисление выхода нечеткой модели*. Выходное значение y определяется на основе вычисленных функций принадлежности к заданным кластерам каждой точки полного координатного пространства интерполируемой поверхности и вычисляется по формуле:

$$y(x_j) = \frac{\sum_{i=1}^{c} m_i^Y \cdot \mu_i^X(x_j)}{\sum_{i=1}^{c} \mu_i^X(x_j)}. \tag{4}$$

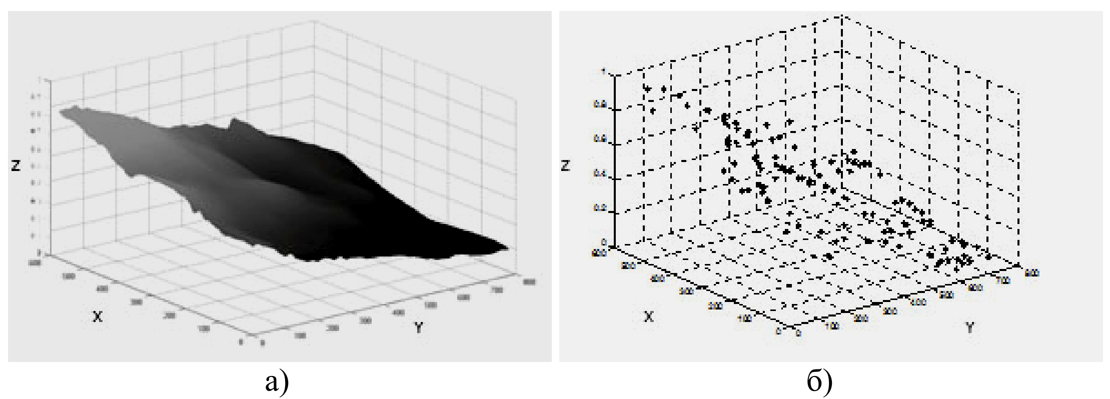
4. *Итеративное уточнение модели*. Полученная модель представляет собой обобщенное представление моделируемой поверхности, поскольку получаемое решение обычно соответствует локальному, а не глобальному максимуму. В методе двумерного проецирования нечетких кластеров повышение точности модели осуществлялось путем формирования дополнительных «существенных» точек на основе анализа ошибки интерполяции в точках координатного пространства из исходного набора данных. В точках пространства, для которых ошибка имела максимальное значение, формировались собственные функции принадлежности на основе выражения (2) и уточнение модели по формуле (4).

Существует множество факторов, влияющих на точность получаемого результата, одним из которых является значение параметра фаззификации. На сегодня не существует теоретически обоснованного правила выбора значения этого параметра.

Значение q влияет на матрицу степеней принадлежности. Чем q больше, тем конечная матрица c-разбиения становится более «размытой». Значение $q = 1$ соответствует «четкой» кластеризации, а при $q \to \infty$ центры кластеров приближаются к точке, соответствующей центру масс всех элементов (все объекты принадлежат ко всем кластерам с одной и той же степенью). При больших значениях q ($q > 1$) усиливается степень нечеткости кластера, т.е. увеличиваются расстояния между ветвями функции принадлежности. Большие значения параметра фаззификации усиливают влияние удаленных элементов, меньшие значения – близких, что особенно важно для кластеризации пространственных данных и позволяет ввести и управлять понятием «ближайшая» окрестность точки.

Экспериментальные результаты

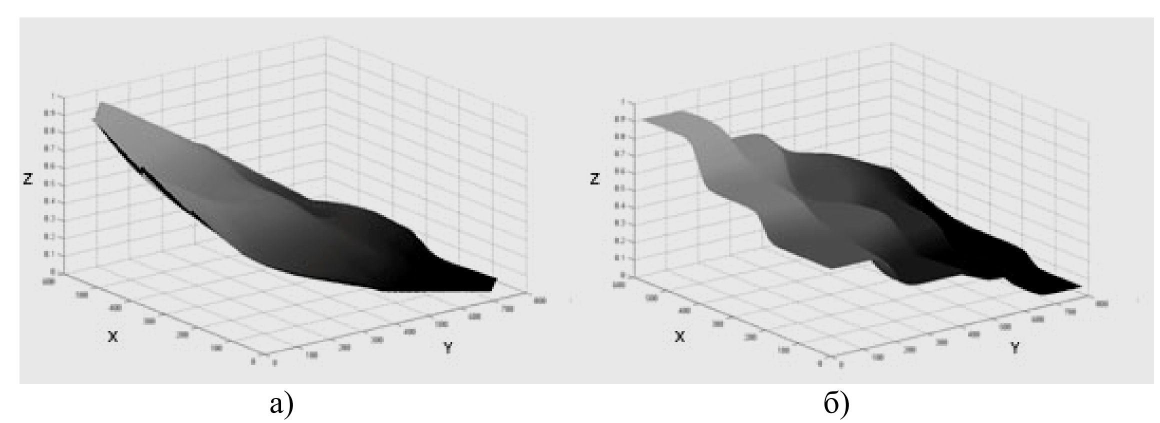
Проверка метода была проведена на модели, в качестве которой рассматривалось геофизическое поле $Z = f(x, y)$, трехмерное изображение которого приведено на рис. 1 а), где (x, y) – координаты поверхности Земли, Z – его нормированное значение, которое может быть интерпретировано как высота поверхности. Формирование исходных пространственных данных, расположенных на неравномерной сетке, которые представлены на рис. 1 б), осуществлялось случайным образом. Количество исходных точек в ходе эксперимента варьировалось в пределах 1 – 5% от общего количества пикселей результирующего изображения поверхности, значение которого определяется шагом дискретизации реальных исходных данных и диапазоном их изменения.



а) б)

Рисунок 1 – Модельные данные: а) – изображение геофизического поля; б) – расположение точек, используемых при интерполяции

На рис. 2 а) представлен результат интерполяции с использованием широко применяемого метода на основе триангуляции Делоне, со сглаживанием участков поверхности кубическими сплайнами, который реализован во многих системах и может быть отнесен к трендовым методам, а на рис. 2 б) – новым методом двумерного проецирования нечетких кластеров.

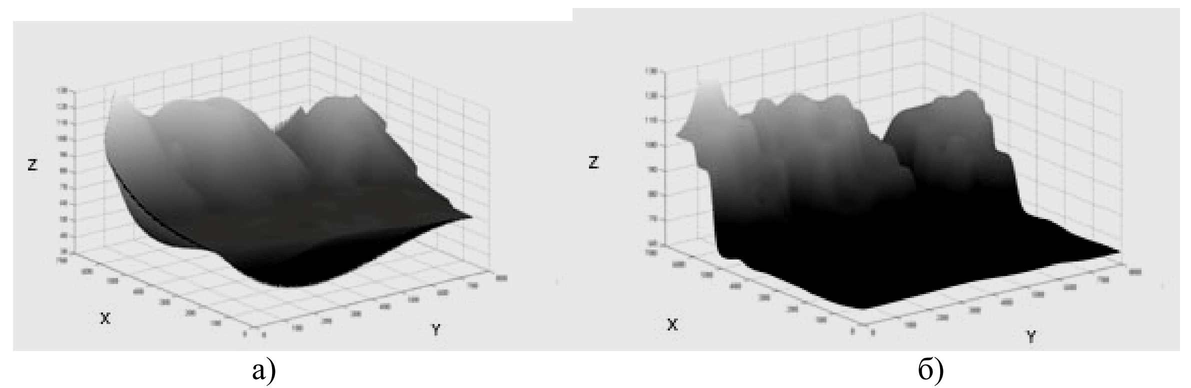


а) б)

Рисунок 2 – Интерполяция модельного поля: а) – методом на основе триангуляции Делоне; б) – методом двумерного проецирования нечетких кластеров

Оценивая полученные результаты, кроме лучшего качественного соответствия модели, следует отметить хороший результат экстраполяции – расширение поверхности на внешние области, для которых отсутствуют результаты измерений (для метода на основе триангуляции Делоне эта область вообще не определена).

На рис. 3 а) – б) представлен результат моделирования участка земной поверхности шахтного поля шахты «Павлоградская» по сведениям из 108 разведочных скважин. Количество исходных данных составляет менее процента от общего количества данных, полученных в результате интерполяции методами: а) – основанным на триангуляции Делоне и б) – новым, двумерного проецирования нечетких кластеров.



а) б)

Рисунок 3 – Интерполяция реальных данных различными методами:
а) – на основе триангуляции Делоне;
б) – двумерного проецирования нечетких кластеров

Выводы

В результате визуального анализа полученных результатов и при сравнении с картографическими источниками было выявлено следующее.

1. Интерполяция неравномерных пространственных данных методом двумерного проецирования нечетких кластеров обеспечивает большее соответствие модели реальной поверхности.
2. Метод обеспечивает лучшее согласование поверхности интерполяции с модельными данными во внешних областях (области экстраполяции).
3. Новый метод имеет значительный потенциал для дальнейшего развития.

Литература

1. Майкл Н. Геоинформационные системы: Основы / Н. Майкл, ДеМерсе. – М., 1999. – 504 с.
2. Пегат А. Нечеткое моделирование и управление / Пегат А. – М. : Бином. – 2009. – 798 с.

Л.Г. Ахметшина, Т.С. Ямнич

Інтерполяція просторових даних методом двовимірного проектування нечітких кластерів

У статті розглядаються інформаційні можливості нового методу інтерполяції просторових даних, заданих на нерівномірній сітці. Суть методу полягає не нечіткій кластеризації неповних експериментальних даних, розгляді проекції функції приналежності на координатну площину, обчисленні ступеня приналежності наявних даних и використання їх для отримання прогнозних значень у додаткових точках. Представлені результати моделювання та реальної перевірки працездатності методу.

L.G. Akhmetshina, T.S. Iamnych

Interpolation of Spatial Data by the Method of Two-dimensional Projection of Unclear Clusters

Informative possibilities of new method of interpolation of spatial data, set on an uneven net are considered. Essence of the method is based on an unclear clustering, consideration of projection of belonging functions on a co-ordinate plane, calculation of present data belonging degree and the use of them for the receipt of prognosis values in additional points. The results of design and real verification of method capacity are presented.

Статья поступила в редакцию 01.07.2010.